АНАЛИЗ ДАННЫХ: ПРОВЕРКА НА НОРМАЛЬНОСТЬ

Цурганов А.Г., Клименок М.Ф.
УО «Витебский государственный ордена Дружбы народов медицинский университет»

Одним из первых вопросов при статистическом анализе медико-биологических данных является проверка соответствия их нормальному закону распределения. Большинство статистических процедур в современных пакетах (t-критерий Стьюдента, ANOVA, корреляционный анализ по Пирсону и др.) ориентированно на выборки, полученные из нормально распределенной генеральной совокупности. Поэтому при использовании параметрических критериев для данных, распределение которых не подчиняется нормальному закону, может привести к неверным выводам.

При каких условиях данные могут быть распределены нормально? Ответ на этот вопрос даёт центральная предельная теорема в форме теоремы Ляпунова Смысл теоремы состоит в том, что сумма n независимых случайных величин, заданных произвольными распределениями, имеет распределение, которое по мере возрастания числа n стремится к нормальному при условии, что влияние каждой случайной величины невелико по сравнению с их суммарным влиянием. Значительное число случайных явлений в природе, технике, медицине протекает именно по такой схеме Например, артериальное давление (АД) зависит от многих факторов: образа жизни, наследственности, погодных условий и т.д., и влияние

многих из них может быть незначительным. Согласно центральной предельной теореме, всегда, когда можно предположить, что рассматриваемая величина (например, АД) является суммой большего числа случайных факторов, влияние каждого из которых мало, её распределение будет близко к нормальному. Для конкретной случайной величины это означает её взаимосвязь со многими подсистемами организма, а не с одной-двумя из них. Такими же нормально распределёнными случайными величинами являются, например, результаты случайного эксперимента, зависящего от многих малых факторов, ошибки регистрации в измерительных приборах и др. Поэтому понимание природы полученных данных облегчает проведение проверки на нормальность.

Можно выделить несколько этапов проверки данных на нормальность. Качественная проверка – это построение гистограммы распределения случайной величины, с помощью которой можно визуально оценить, насколько гистограмма близка к колоколу нормального распределения. Поскольку вид гистограммы очень чувствителен к количеству интервалов, на которые разбивается диапазон случайных величин, рекомендуется определять количество интервалов по формуле Стерджеса, кроме того, необходимо учитывать, что для любой симметричной выборки гистограмма будет иметь скос влево или вправо, что теоретически обосновано.

Вторым визуальным способом проверки на нормальность является построение пробит-графиков, т.е. графиков функций, обратных к некоторой функции распределения, например, нормального распределения. По степени отклонения пробит-графика от прямой линии судят о близости распределения к нормальному, а также о наличии выбросов.

Визуальные методы на основе гистограмм и пробит-графиков могут дать только качественные первоначальные предположения о распределении. Более надежны выводы на основе числовых характеристик выборки. Во-первых, вычислив предварительно S данной выборки, можно проверить «правило трех сигм», согласно которому практически достоверно (p=99,7%), что нормально распределенная случайная величина попадает в интервал ± 3S в единичном испытании.

При большом числе наблюдений ($n > 100$) хорошие результаты даёт вычисление параметров формы распределения – ассиметрии и эксцесса и их стандартных ошибок Коэффициент ассиметрии As – это характеристика скошенности распределения вправо или влево относительно максимума Для симметричных распределений показатель ассиметрии равен нулю: $As = 0$ Если $As < 0$, то кривая распределения скошена влево относительно симметричной кривой (левый хвост длиннее правого), а если $As > 0$, то кривая скошена вправо (правый хвост частотной гистограммы длиннее левого). Распределение считается симметричным, если $|As| \leq 0,1$ и ассиметричным, если $|As| > 0,5$. Показатель эксцесса Ex – это характеристика плосковершинности кривой распределения. Для нормальной кривой показатель эксцесса равен нулю: $Ex=0$ Распределение близко к нормальному, если $|Ex| \leq 0,1$ и значительно отклоняется от него, если $|Ex| > 0,5$ При этом найденные стандартные ошибки As и Ex должны иметь тот же порядок, что и сами значения As и Ex.

Наиболее убедительные результаты при проверке нормальности даёт использование критериев согласия – статистических критериев, предназначенных для проверки согласия опытных данных и теоретической модели Выдвигается нулевая гипотеза H_0: выборка получена из нормальной генеральной совокупности

и альтернативная гипотеза H_1: распределение генеральной совокупности отличается от нормального. Далее рассчитывается уровень значимости, соответствующий полученному значению статистики критерия: если $p > 0,05$, то нулевая гипотеза о нормальности выборки принимается; если $p < 0,05$, то H_0 отклоняется и соответственно принимается H_1.

Наиболее часто используют следующие критерии:

1) Хи – квадрат (χ^2) для дискретных и непрерывных величин;
2) W – тест Шапиро-Уилка для малых выборок; среднее значение и среднее квадратическое значения выборки заранее неизвестны и вычисляются по выборке;
3) Лиллиефорса, среднее значение и среднее квадратическое значения выборки заранее неизвестны;
4) Колмогорова-Смирнова для непрерывных (реже дискретных) величин; более мощный, чем χ^2 Среднее значение и среднее квадратическое значения выборки известны заранее;
5) ω^2 Мизеса для непрерывных величин.

Критерии согласия имеют ограничения по объёму выборки.

- для критерия χ^2: $n > 30$;
- для критериев ω^2, Колмогорова-Смирнова: $n > 50$.

Очень часто в медико-биологических исследованиях используется малые выборки, в этом случае рекомендуется для проверки нормальности использовать критерий Шапиро-Уилка, предназначенный для выборок с численностью от 3 до 50 наблюдений. Данный критерий предпочтителен, т.к. является наиболее мощным и универсальным и при этом наиболее строгим из перечисленных выше Если гипотеза о нормальном распределении отклонена, но имеются выпадающие значения, то после их удаления необходимо ещё раз провести тест на нормальность. По данным В.П. Леонова и П.В. Ижевского (1995), лишь около 20% выборочных распределений, встречающихся в медико-биологических исследованиях, являются приближенно нормальными. Если распределение в выборке можно признать нормальным, то лучше использовать параметрические методы, которые мощнее непараметрических, т.е. чаще, чем непараметрические обнаруживают существующие различия выборок.